Bioinformatique BTV
Alignement de Séquences

Jean-Michel Richer
jean-michel.richer@univ-angers.fr
http://www.info.univ-angers.fr/pub/richer

FACULTÉ
DES SCIENCES
Unité de formation
et de recherche

Juillet 2008

Plan

1. Rappels
2. Partie théorique
   - Alignement par paires
   - Alignement multiple
3. Partie pratique

Rappels

Rappels

Biologie moléculaire

Definition (Biologie Moléculaire - Warren Weaver, 1938)

La biologie moléculaire (où *bio mol*) est une discipline scientifique qui vise à comprendre les mécanismes de fonctionnement de la cellule au niveau moléculaire

Historique

- 1930 : techniques de diffraction à rayons X
- 1953 : découverte de la structure de l'ADN par Watson et Crick
- 1977 : séquençage de l'ADN par Gilbert et Sanger
- 2004 : séquençage du génome humain (HUGO)

Biologie moléculaire

Definition (Biologie Moléculaire - Warren Weaver, 1938)

La biologie moléculaire (où *bio mol*) est une discipline scientifique qui vise à comprendre les mécanismes de fonctionnement de la cellule au niveau moléculaire

Historique

- 1930 : techniques de diffraction à rayons X
- 1953 : découverte de la structure de l'ADN par Watson et Crick
- 1977 : séquençage de l'ADN par Gilbert et Sanger
- 2004 : séquençage du génome humain (HUGO)

Evolution du point de vue moléculaire

Modifications

- *point mutation* : modification d'un AN ou AA
- *insertion* : ajout d'un nouvel AN ou AA
- *deletion* : suppresion d'un AN ou AA
- recombinaison des gènes

Résultats

- mauvais repliement $\Longrightarrow$ fonction ineffective
- apparition d'une nouvelle fonction ($\Longrightarrow$ nouvelle espèce ?)

Evolution du point de vue moléculaire

Modifications

- *point mutation* : modification d'un AN ou AA
- *insertion* : ajout d'un nouvel AN ou AA
- *deletion* : suppresion d'un AN ou AA
- recombinaison des gènes

Résultats

- mauvais repliement $\Longrightarrow$ fonction ineffective
- apparition d'une nouvelle fonction ($\Longrightarrow$ nouvelle espèce ?)

Dogme central

Dogmes lié à l'alignement

- les AN ou AA essentiels à la fonctions sont moins sujets à mutation
- plus deux séquences se ressemblent, plus elles ont une forte probabilité de se comporter de manière identique.

Un alignement permet l'identification

- de motifs fonctionnels ou structurels conservés
- de zones non conservées qui résultent d'évènements spécifiques

Dogme central

Dogmes lié à l'alignement

- les AN ou AA essentiels à la fonctions sont moins sujets à mutation
- plus deux séquences se ressemblent, plus elles ont une forte probabilité de se comporter de manière identique.

Un alignement permet l'identification

- de motifs fonctionnels ou structurels conservés
- de zones non conservées qui résultent d'évènements spécifiques

Objectif de l'alignement

Definition (Alignement)

En bioinformatique, l'opération d'alignement vise à identifier des zones communes à un groupe de k séquences.

Definition (Similarité et homologie)

des zones qui se ressemblent sont dites similaires ou homologues si elles dérivent d'un ancêtre commun

Objectif de l'alignement

Definition (Alignement)

En bioinformatique, l'opération d'alignement vise à identifier des zones communes à un groupe de k séquences.

Definition (Similarité et homologie)

des zones qui se ressemblent sont dites similaires ou homologues si elles dérivent d'un ancêtre commun

Applications de l'alignement

Applications

- étude phylogénétique
- étude comparative des génomes (comparative genomics)
- prédiction de gène
- prédiction de la structure 2D/3D des protéines
- caractérisation de la fonction des protéines
- prédiction de la structure et fonction des ARN
- réseaux d'interaction
- génétique (différence entre génotype et phénotype)
- découverte et conception de médicaments

Partie théorique

Partie théorique

Types d'alignements

Definition (Alignement local ou global)

- global : on tente d'identifier des similarités sur la longueur totale des séquences ($\rightarrow$ pb si séquences de longueur différentes)
- local : on tente d'identifier des similarités entre une séquence et une sous-séquence

Definition (Alignement par paires ou multiple)

- par paires : on aligne 2 séquences
- multiple : on aligne plus de 2 séquences

Types d'alignements

Definition (Alignement local ou global)

- global : on tente d'identifier des similarités sur la longueur totale des séquences ($\rightarrow$ pb si séquences de longueur différentes)
- local : on tente d'identifier des similarités entre une séquence et une sous-séquence

Definition (Alignement par paires ou multiple)

- par paires : on aligne 2 séquences
- multiple : on aligne plus de 2 séquences

Alphabet

Definition (Alphabet)

Un alphabet $\Sigma = \{a_0, a_1, \ldots, a_n\}$ est un ensemble fini de symboles distincts deux à deux. En particulier, le symbole a_0 est appelé *brêche* ou *gap* (en anglais) et est représenté par le caractère $-$.

Par la suite, nous utiliserons de manière préférentielle le terme gap plutôt que le terme brêche.

Alphabets biologiques

Definition (Alphabet de l'ADN)

L'alphabet des molécules d'ADN est composé de 5 symboles $\Sigma_{ADN} = \{-, A, C, G, T\}$ qui représentent respectivement un gap, l'Adénine, la Cytosine, la Guanine et la Thymine.

Definition (Alphabet de l'ARN)

L'alphabet des molécules d'ARN est composé de 5 symboles $\Sigma_{ARN} = \{-, A, C, G, U\}$ qui représentent respectivement un gap, l'Adénine, la Cytosine, la Guanine et l'Uracile.

Definition (Alphabet des Protéines)

L'alphabet des protéines est composé de 21 symboles $\Sigma_{AA} = \{-, A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ qui représentent les différents acides aminés.

Séquence et sous-séquence

Definition (Séquence)

On appelle séquence S une suite ordonnée de caractères $S = \langle x_1, x_2, \ldots, x_n \rangle$ pris dans un alphabet Σ. On note $|S| = n$ la longueur de la séquence.

Definition (Sous-séquence)

Soit S une séquence de longueur n. On appelle sous-séquence de S toute partie de S composée d'un ensemble de caractères consécutifs de S. On notera $S[i..j]$ avec $1 \leq i \leq j \leq n$, la sous-séquence $\langle x_i, x_{i+1}, \ldots, x_j \rangle$. En particulier $S[i..i] = S[i] = \langle x_i \rangle$.

Distance

Definition (Métrique ou distance)

On rappelle qu'une métrique sur un ensemble X est une application $d : X \times X \to \mathbb{R}$ vérifiant les propriétés suivantes :

- $d(x, y) \geq 0$, non négativité,
- $d(x, y) = 0 \iff x = y$, identité des indiscernables,
- $d(x, y) = d(y, x)$, symétrie,
- $d(x, z) \leq d(x, y) + d(y, z)$, inégalité triangulaire.

Distance de Hamming

Definition (Distance de Hamming)

Soient deux séquences de même longueur S et T, la distance de Hamming de S et T, notée $d_H(S, T)$, correspond au nombre de caractères en regard qui diffèrent. Plus la distance de Hamming est faible, plus les séquences sont proches.

Distance de Hamming

Example

Si l'on considère les séquences suivantes :

$$\begin{array}{ll} S_1 & \text{ACACACACACAT} \\ S_2 & \text{ACACAGACATAT} \\ S_3 & \text{CACACACACATA} \end{array}$$

$$\begin{aligned} d_H(S_1, S_2) &= 2 \\ d_H(S_1, S_3) &= 12 \\ d_H(S_2, S_3) &= 12 \end{aligned}$$

Opérations d'édition

Definition (Opération d'édition d'une séquence)

- *l'appariement* ou mouvement diagonal qui consiste à placer en regard des caractères qui ne sont pas des gaps de manière à faire appraître :
  - soit des *conservations* pour lesquelles les caractères en regard sont égaux (a, a),
  - soit des *substitutions* pour lesquelles les caractères en regard sont différents (a, b), a dans S est en regard de b dans T. Il s'agit ici de faire apparaître une possible mutation ou d'éviter d'introduire un nombre trop important de gaps,
- *l'insertion d'un gap* dans S : $(-, b)$, nous qualifierons ce mouvement de vertical,
- *l'insertion d'un gap* dans T : $(a, -)$, mouvement horizontal

Opérations d'édition

Example

Séquences $S = \langle$ CATGC $\rangle$ et $T = \langle$ ACAGTC $\rangle$:

```
-CA-TGC
ACAGT-C
```

S	T	Opération	Description
-	A	(-,A)	insertion de $-$ dans S
C	C	(C,C)	appariement sur C
A	A	(A,A)	appariement sur A
-	G	(-,G)	insertion de $-$ dans S
T	T	(T,T)	appariement sur T
G	C	(G,C)	substitution de G par C
C	-	(C,-)	insertion de $-$ dans T

Distance d'édition

Definition (Distance d'édition)

A partir des opérations nécessaires pour obtenir un alignement, on peut calculer une distance dite distance d'édition ou de Levenshtein [5], définie par :

$$d_L(S,T) = \sum_{i=1}^{q} d(x_i, y_i) = \begin{cases} 0 & \text{si} x_i = y_i \\ 1 & \text{sinon} \end{cases}$$

Example

Dans l'exemple précédent, la distance d'édition est de 4 et correspond à trois insertions et une substitution.

Alignement

Definition (Alignement (1/2))

Soit $S = \{S_1, \ldots, S_k\}$ un ensemble de k séquences définies sur un alphabet Σ : $\forall u, 1 \leq u \leq k, S_u = \langle x_1^u, \ldots, x_{|S_u|}^u \rangle$. Un alignement $A(\{S_1, \ldots, S_k\})$ est une matrice :

$$A = \begin{bmatrix} a_1^1 & \ldots & a_q^1 \\ \vdots & & \vdots \\ a_1^k & & q_q^k \end{bmatrix}$$

avec $\forall u \in \{1, \ldots, k\}, \forall v \in \{1, \ldots, q\}, \quad a_v^u \in \Sigma$.

Alignement

Definition (Alignement (2/2))

La matrice A vérifie les propriétés suivantes :

1. **cohérence sur la longueur :**

$$\forall u \in \{1, \ldots, k\}, \quad max(|S_u|) \leq q \leq \sum_{u=1}^{u=k} |S_u|$$

2. **absence de colonne de gaps :**
$\nexists j \in \{1, \ldots, q\}$ tel que $\forall u \in \{1, \ldots, k\}, \quad a_j^u = -$
3. **conservation des séquences initiales :** pour tout $u \in \{1, \ldots, k\}$, il existe un isomorphisme d'ordre $f_u : \{1, \ldots, |S_u|\} \rightarrow \{1, \ldots, q\}$ tel que $\langle a^u_{f_u(1)}, a^u_{f_u(2)}, \ldots, a^u_{f_u(|S_u|)} \rangle = S_u$

Matrice de substitution

Definition (Matrice de substitution)

Une Matrice de substitution permet d'attribuer un score aux opérations d'appariement (conservation ou substitution). Une Matrice de substitution est donc une application w définie sur un alphabet $\Sigma = \{a_0, a_1, \ldots, a_n\}$ telle que $w : \Sigma \times \Sigma \rightarrow \mathbb{R}$. Nous imposons que w vérifie les propriétés suivantes :

$$w(a_0, a_0) = 0$$
$$w(a_0, x) = w(x, a_0) = 1, \forall x \neq a_0 \in \Sigma$$

Matrice de substitution

Example

Scores Soit la matrice de substitution $w(x,y)$ pour l'alphabet $\Sigma = \{-, A, C, G, T\}$ donnée par :

$$w(x,y) = \begin{bmatrix} x/y & - & A & C & G & T \\ - & 0 & 1 & 1 & 1 & 1 \\ A & 1 & 6 & 2 & 2 & 2 \\ C & 1 & 2 & 6 & 2 & 2 \\ G & 1 & 2 & 2 & 6 & 2 \\ T & 1 & 2 & 2 & 2 & 6 \end{bmatrix}$$

Les scores $w(x,y)$ des opérations d'alignement sont donnés par :

- 6, s'il s'agit d'un appariement,
- 2, s'il s'agit d'une substitution,

Matrices de substitution

Matrices liées aux AA

- PAM (*Point Accepted Mutation* - Henikoff [1])
- BLOSUM (*BLOck SUbstitution Matrices* - Dayhoff [4])
- Gonnet [2]

Relations entre matrices

- séquences peu divergentes : BLOSUM80, PAM1
- séquences très divergentes : BLOSUM45, PAM250
- en général : BLOSUM62, PAM120
- séquences courtes PAM30 ($<$ 35 AA), PAM70 ($<$ 50 AA)

Modèle de gap

Definition (Modèle de gap)

Un modèle de gap est une application g de $\mathbb{N} \rightarrow \mathbb{R}$ qui attribue un score, qualifié ici de pénalité, à un ensemble de gaps consécutifs. Cette pénalité possède un score généralement négatif.

Definition (Modèle de gap linéaire)

Dans ce modèle, le score d'un gap est proportionnel à la longueur du gap et est donné par une formule de la forme :

$$g(n) = \begin{cases} 0 & \text{si} \quad n = 0 \\ n \times g_o & \text{si} \quad n \geq 1 \end{cases}$$

ou $g_o < 0$ est la pénalité introduite par l'insertion d'un nouveau gap et n est le nombre de caractères gap consécutifs.

Modèle de gap

Definition (Modèle de gap affine)

La fonction de score est donnée par :

$$g(n) = \begin{cases} 0 & \text{si} \quad n = 0 \\ g_o + (n-1) \times g_e & \text{si} \quad n \geq 1 \end{cases}$$

ou $g_o < 0$ est la pénalité d'introduction (*gap opening penalty*) d'un nouveau gap et $g_e < 0$ est la pénalité d'extention d'un gap existant (*gap extension penalty*).

Somme des paires d'un alignement

Definition (Somme des paires d'un alignement)

Soit A un alignement, la somme des paires (ou score de l'alignement) est donnée par la formule :

$$sop(A) = \sum_{c=1}^{q} sop^c(A_c)$$

ou $sop^c(A_c)$ est le score de la colonne c de l'alignement, donné par :

$$sop^c(A_c) = \sum_{r=1}^{k-1} \sum_{s=r+1}^{k} \delta_{r,s} \times w(a_c^r, a_c^s)$$

où $0 < \delta_{r,s} \leq 1$ est un coefficient de pondération

Meilleur alignement par paires

Comment calculer le meilleur alignement ?

On utilise

- une matrice de substitution
- un modèle de gap
- une fonction de score (somme des paires)

Le meilleur alignement est l'alignement optimum pour la somme des paires

Meilleur alignement par paires

Comment obtenir le meilleur alignement ?

- énumération exhaustive : $\sum_{k=0}^{n} C_{n+k}^{k} \times C_{n}^{k}$
- méthode heuristique
- méthode exacte : programmation dynamique

Programmation Dynamique

Programmation dynamique (Bellman, 1940)

méthode appliquée à des problèmes d'optimisation pour lesquels un choix doit être fait entre plusieurs solutions possibles afin d'aboutir à une solution optimale. Le terme Programmation fait ici référence à une méthode basée sur le calcul de tableaux de valeurs (Needleman et Wunsch, 1970 [6])

Complexité

en $O(n \times p)$ si séquences de longueurs respectives n et p.

Programmation Dynamique

Principe - cas d'un gap linéaire

- soient 2 séquences S et T à aligner de longueurs N et P
- $S = \langle x_1, \ldots, x_N \rangle$
- $T = \langle y_1, \ldots, y_P \rangle$
- on calcule une matrice M de scores optimaux de dimension $(N + 1) \times (P + 1)$
- à partir de cette matrice on peut évaluer les alignements optimaux

Programmation Dynamique

Initialisation la matrice M

- $M[0,0] = 0$
- $M[i,0] = M[i-1,0] + g_o \quad \forall i \in [1,N]$
- $M[0,j] = M[0,j-1] + g_o \quad \forall j \in [1,P]$

Programmation Dynamique

Calcul de chaque case de la matrice

$$\begin{array}{ccc} M[i-1,j-1] & & M[i-1,j] \\ & \nwarrow \quad \uparrow & \\ M[i,j-1] & \leftarrow & M[i,j] \end{array}$$

Formule de récurrence

$$M[i,j] = max \begin{cases} M[i-1,j-1] & +w(x_i,y_j) \\ M[i-1,j] & +g_o \\ M[i,j-1] & +g_o \end{cases}$$

Programmation Dynamique

Example

Exemple 1

- $S = ACAGTC$
- $T = CATTGC$
- $w(a, a) = 1$
- $w(a, b) = 0$
- $g_o = 0$

Programmation Dynamique

Exemple 1 - initialisation

S/T	-	A	C	A	G	T	C	i
-	0	0	0	0	0	0	0	0
C	0							1
A	0							2
T	0							3
T	0							4
G	0							5
C	0							6
j	0	1	2	3	4	5	6	

Programmation Dynamique

Exemple 1 - calcul de *M*

S/T	-	A	C	A	G	T	C	i
-	0	0	0	0	0	0	0	0
C	0	0	1	1	1	1	1	1
A	0	1	1	2	2	2	2	2
T	0	1	1	2	2	2	2	3
T	0	1	1	2	2	3	3	4
G	0	1	1	2	3	3	3	5
C	0	1	2	2	3	3	**4**	6
j	0	1	2	3	4	5	6	

Programmation Dynamique

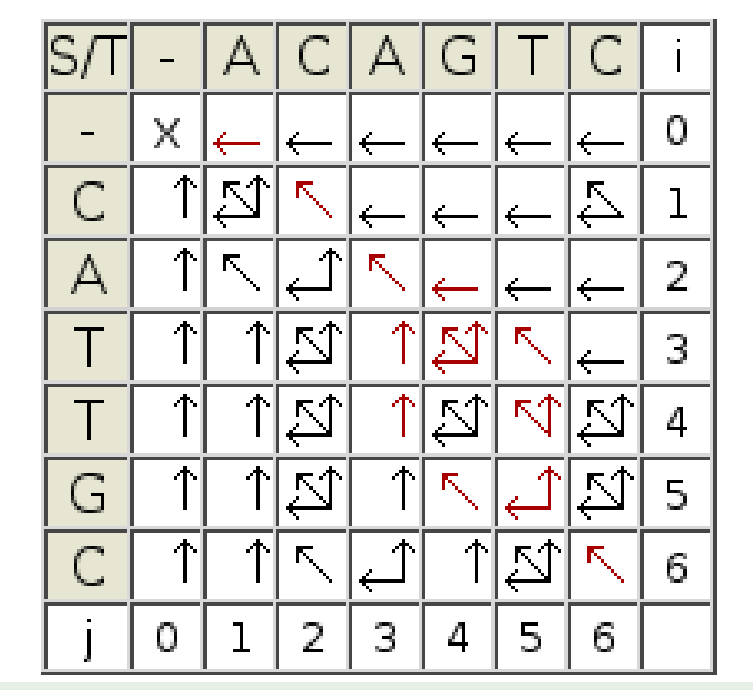
Obtention de l'alignement

A partir de la matrice des scores optimaux M on obtient les alignements comme suit :

- on part de la case $M[N, P]$
- on prend une direction qui correspond au calcul optimal

Programmation Dynamique

Exemple 1 - directions en fonction de *M*



Programmation Dynamique

Exemple 1 - Alignements

On obtient 5 alignements optimaux :

-CATTG-C ACA--GTC	-CAT-TGC ACA-GT-C	-CATTGC ACAGT-C	-CA-TTGC ACAG-T-C

-CA-TTGC ACAGT--C

Programmation Dynamique

Autres types d'alignement

le même principe peut être appliqué :

- à l'alignement global avec gap affine (Gotoh 82 [3])
- à l'alignement local (Smith et Waterman 81 [7])

Autres méthodes - BLAST, FASTA

Recherche dans les bases de données

Lorsque l'on doit réaliser de très nombreux alignements, l'algorithme de programmation dynamique est trop coûteux.
Deux algorithmes heuristiques ont été développés :

- BLAST
- FASTA

Méthodes d'alignement multiple

Programmation dynamique à k-dimensions

- on peut étendre l'algorithme de programmation dynamique pour trouver l'alignement optimal de k séquences. Cependant, cet algorithme est trop coûteux en espace mémoire et en temps pour être efficace
- il est donc nécessaire de développer des algorithmes sous-optimaux mais efficaces

Méthodes d'alignement multiple

Progressif ou itératif

On distingue 2 grand types de méthodes :

- **progressives** (Clustal) : on commence par aligner les deux séquences les plus proches, puis on ajoute les séquences de plus en plus distantes au fur et à mesure
- **itératives** (Saga) : on aligne l'ensemble des séquences et on améliore l'alignement par une série d'étapes

Remarque

Les algorithmes progressifs sont plus rapides que les algorithmes itératifs.

Les programmes d'alignement multiple

- `clustalw` : progressif
- `multalin` : variante de clustal
- `T-coffee` : variante de clustal
- `muscle` : fonction de création de profile
- `probcons` : modèle de Markov
- `mafft` : transformée de Fourier
- `dialign` : recherche de chemins
- `saga` : algorithme génétique
- `hmmer` : modèle de Markov
- ...

Utilitaires pour l'alignement

- `readseq` : conversion entre différents format de séquences
- `cinema` : visualisation d'alignement multiple

Efficacité et précision

BaliBase (Thompson, Plewniak, Poch 99 [8])

- ensemble d'alignements de référence (considérés corrects)
- utilisé pour attester de la qualité des logiciels d'alignement multiple
- décomposé en 5 sous-ensembles caractéristiques :
  - set 1 : séquences équidistantes
  - set 2 : une séquence orpheline
  - set 3 : familles divergentes
  - set 4 : longues insertions de gap aux extrémités
  - set 5 : longues insertions de gap au milieu

Efficacité et précision

BaliBase (bali_score)

Le programme `bali_score` permet de calculer 2 valeurs :

- SPS (sum-of-pairs score) : rapport entre le nombre de paires de résidus bien alignés dans l'alignement obtenu et ceux de l'alignement de référence
- CS (column score) : nombre de colonnes bien alignées par rapport au nombre de colonnes de l'alignement

de manière générale

$0 \leq CS \leq SPS \leq 1$

Efficacité et précision

Résultats du score SPS avec quelques logiciels :

Softwares	Set 1	Set 2	Set 3	Set 4	Set 5	Time (s)
CLUSTAL	0.809	0.932	0.723	0.834	0.858	120
MAFFT	0.829	0.931	0.812	**0.947**	**0.978**	98
MUSCLE	0.821	0.935	0.784	0.841	0.972	75
PROBCONS	**0.849**	**0.943**	**0.817**	0.939	0.974	711
TCOFFEE	0.814	0.928	0.739	0.852	0.943	1653
MALINBA	0.811	0.911	0.752	0.899	0.942	343

Partie pratique

Partie pratique

Installer Clustalw

Mode d'emploi

- télécharger clustalw 1.83
- désarchiver le fichier : `tar -xvzf *.tgz`
- compiler : `make -f makefile.linux`

Utiliser Clustalw

Mode interactif ou non

On peut utiliser Clustal de deux manières différentes :

- soit de manière interactive (l'utilisateur saisit au clavier les différents paramètres) : `clustalw`
- soit de manière non-interactive (on fournit les paramètres en ligne de commande)

pour connaître les paramètres en ligne de commande :
`clustalw -help`

Utiliser Clustalw

Alignement de séquences

Avec clustalw, aligner les séquences des fichiers :

- `1aab_ref1.fasta`
- `1aho_ref1.fasta`
- `1csy_ref1.fasta`
- `1dox_ref1.fasta`

puis calculez le SPS et le CS de des alignements obtenus

M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt.
A model of evolutionary change in proteins.
In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, chapter 22, pages 345–352. National Biomedical Research Foundation, 1978.

G.H. Gonnet, M.A. Cohen, and S.A. Benner.
Exhaustive matching of the entire protein sequence database.
Science, 256 :1443–1445, 1992.

O. Gotoh.
An improved algorithm for matching biological sequences.
Journal of Molecular Biology, Vol. 162 :705–708, 1982.

S. Henikoff and J. G. Henikoff.
Amino acid substitution matrices from protein blocks.

In *Proceedings of the National Academy of Science*, volume Vol. 89, pages 10915–10919, 1992.

V. I. Levenshtein.
Binary codes capable of correcting deletions, insertions, and reversals.
Soviet Physics - Doklady, 10(8) :707–710, February 1966.

Wunsch C.D. Needleman S.B.
A general method applicable to the search for similarities in the amino acid sequence of two proteins.
JMB, 3(48) :443–453, 1970.

T. F. Smith and M. S. Waterman.
Identification of common molecular sequences.
JMB, 147 :195–197, 1981.

J.D. Thompson, F. Plewniak, and O. Poch.

Balibase : A benchmark alignments database for the evaluation of multiple sequence alignment programs. *Bioinformatics*, Vol. 15 :87–88, 1999.